PURA grant proposal (Fall 2017)
Computational Painting of Local Genetic Ancestry Along Chromosomes
Greg Johnston (Mentor: Dr. Joseph Lachance)

Background

Each of our genomes is a mosaic of chromosomal fragments inherited from different ancestries. These fragments are not necessarily spread evenly across any of our genomes. Rather, their size and distribution can vary between populations and even between individuals in the same population, and can be a useful source of information for identifying admixture trends and more. The task of characterizing this mixture – of identifying the ancestral population of origin at each position in an organism's genome – is referred to as local ancestry inference. This information can be used to understand when admixture has occurred in a population (recent mixing of populations leads to large ancestry fragments) and identify likely targets of natural selection (adaptive parts of the genome can be enriched for a particular ancestry). Local ancestry inference is one service offered by commercial genomics companies like 23andMe and AncestryDNA (Figure 1).

Multiple approaches have been developed for the process of local ancestry inference. LAMP draws inference across more than two ancestral populations at once, and combines a hidden Markov model with a framework of non-overlapping windows to achieve higher accuracy and speed.[1] RFMix combines information from reference individuals and admixed samples and performs inference even more quickly and accurately than LAMP, with the ability to distinguish between subcontinental ancestries.[2]

A disadvantage common to most current algorithms for local ancestry inference (including the two described above) is that they require a large set of pure reference individuals for each ancestral population that they can identify. This can be problematic since some populations of interest are now extinct, with limited genetic data available. Even if a population is still present today, obtaining enough sequenced data may not be possible.

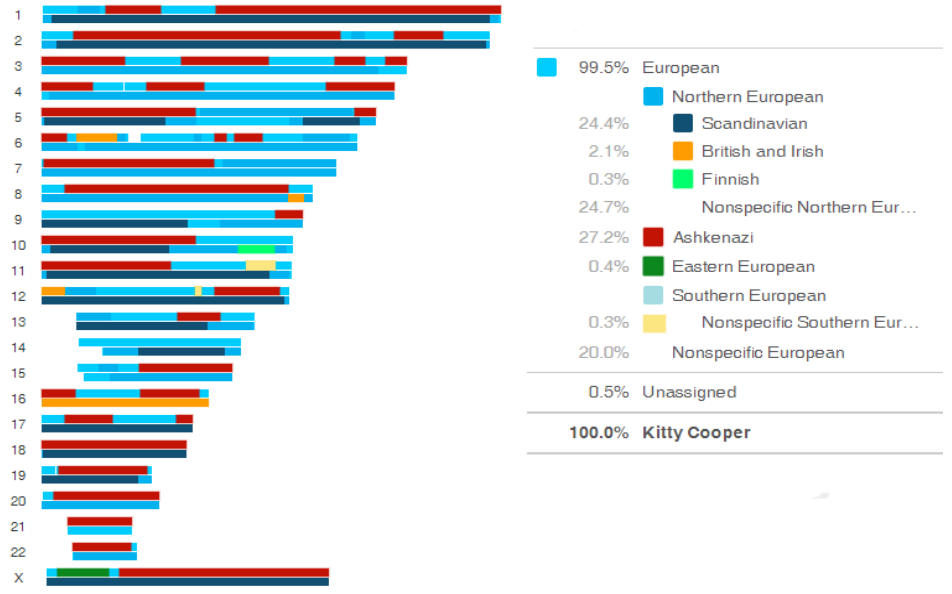

Figure 1: Local ancestry painting summary from 23andMe[3]

Previous research (STRUCTUREpainter)

STRUCTUREpainter is a novel approach for performing local ancestry inference, currently under development at the Lachance Lab. It has the distinct advantage of not requiring reference panels for each ancestral population of interest - it only requires a set of admixed individuals and their estimated ancestry proportions. The algorithm was originally implemented in R by Dr. Ali Berens working with Dr. Lachance. My work so far has focused on reimplementing it in Python, leading to improved flexibility and a speedup of over 100%.

STRUCTUREpainter works by finding SNPs (single-nucleotide polymorphisms) that are associated with one ancestry or another, by analyzing the provided set of admixed individuals and estimated ancestry proportions. The probability of transitions between ancestries is also calculated, based on the overall ancestry proportions and a fixed recombination rate. Then, given an individual of unknown ancestry, ancestry for each chromosome is estimated sequentially (starting at one end of the chromosome and working towards the other end), based on transition probabilities as well as the relative probabilities that a SNP was inherited from one ancestral population or another.

Figure 2 shows output from STRUCTUREpainter on a test chromosome composed of African and European ancestries. Large ancestry fragments are mostly correctly identified, with some incorrectly identified small fragments.

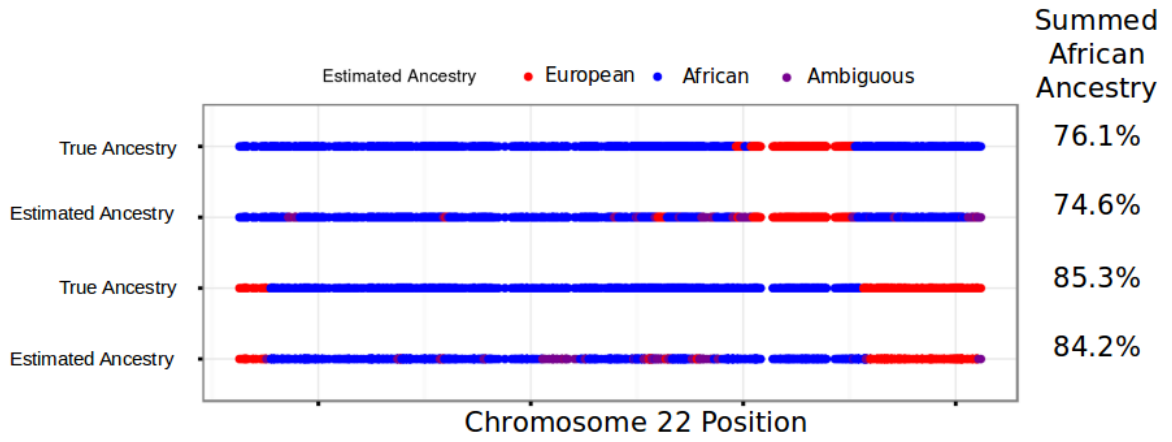

Figure 2: STRUCTUREpainter yields accurate inference of local ancestry fragments for African genomes

Future research plans (Fall 2017 semester)

Although STRUCTUREpainter is completely functional in its current state, there are still ways in which it can be improved. In the Fall 2017 semester, my research in the Lachance Lab will focus on the following:

- ***Aim 1: improve the accuracy of STRUCTUREpainter***
  - The number of SNPs to which ancestry is assigned at each step in the algorithm is an important parameter. In practice, the optimal value for this parameter may vary based on the testing set, so further investigation into how to determine this parameter in each instance will be useful.
  - Using different statistics to determine how helpful each SNP is for indicating ancestry is another modification which could lead to substantial improvements in accuracy.
  - Post-processing of the results to resolve ambiguous ancestry calls is an important step that will almost certainly improve accuracy.

- ***Aim 2: improve the computational speed of STRUCTUREpainter***
  - Adding the capability of parallelization (simultaneously computing ancestry for multiple chromosomes) to the program would be extremely helpful, since practical applications of STRUCTUREpainter will most likely involve running it on each chromosome for each individual in a large population.

- ***Aim 3: apply STRUCTUREpainter to a novel dataset***
  - Sequenced data has recently become available for the Norfolk Island population. This population of mixed European and Polynesian ancestry is descended from the survivors of the 1789 mutiny on the HMS Bounty. The combination of the founder effect and the population's isolation and unique ancestry makes this a unique opportunity for study. Knowledge of local ancestry would serve many purposes, and since modern Polynesian reference genomes are unavailable, STRUCTUREpainter is uniquely equipped for this task.

Personal benefits

In addition to the practical applications described above, this project has allowed me to gain and develop many technical skills: I was exposed to R for the first time in the process of interpreting and translating the existing R code, and I have been improving my proficiency with Python scientific computing and data manipulation/visualization libraries. I have also gained experience using genetics software tools such as admixture, plink, and vcftools. This has been my first experience with bioinformatics, and seeing how the concepts and skills that I have learned as a computer science major can be applied in the context of biology has been interesting and rewarding. Mostly due to my experience with this project, I am now heavily considering a graduate degree in bioinformatics.

Works Cited:

1) Y. Baran, B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux, C. Eng, et al. "Fast and Accurate Inference of Local Ancestry in Latino Populations." *Bioinformatics* 28.10 (2012): 1359-367.
2) B. Maples, S. Gravel, E. Kenny, C. Bustamante. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *American Journal of Human Genetics* 93.2 (2013): 278-288.
3) K. Cooper. “23andMe ancestry painting summary” (2012). Retrieved from http://blog.kittycooper.com/2012/12/new-ancestry-composition-tool-at-23andme/